Taxonomies for Text Analytics and Auto-Indexing

Heather Hedden
Hedden Information Management

Text Analytics World, Boston, MA

October 4, 2012

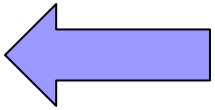
Introduction: Text Analytics and Taxonomies

- Text analytics can be used to index content without the use of taxonomies/controlled vocabularies.

- Text analytics can be used to index content *with* taxonomies/controlled vocabularies for better results.

Text analytics can generate terms from text to be used:

1. As a source to manually build taxonomies
2. To auto-categorize/classify content against existing taxonomies



Outline

- Taxonomy Introduction: Definitions & Types
- Taxonomy Introduction: Purposes & Benefits
- Synonyms for Terms
- Auto-Indexing and Auto-Categorization
- Taxonomies for Auto-Categorization
- Taxonomy Resources

Outline

- Taxonomy Introduction: Definitions & Types
- Taxonomy Introduction: Purposes & Benefits
- Synonyms for Terms
- Auto-Indexing and Auto-Categorization
- Taxonomies for Auto-Categorization
- Taxonomy Resources

Definitions & Types

Broad designations

(essentially the same meaning – used interchangeably):

- Controlled Vocabularies (CV)
- Knowledge Organization Systems
- Taxonomies

Specific types

(different meanings):

- Term Lists/Pick lists
- Synonym Rings
- Authority Files
- Taxonomies
  - Hierarchical
  - Faceted
- Thesauri
- Ontologies

Definitions & Types

Broad Designations:

Controlled vocabulary, knowledge organization system, taxonomy

- An authoritative, restricted list of terms (words or phrases)
- Each term for a single unambiguous concept (synonyms/nonpreferred terms, as cross-references, may be included)
- Policies (control) for who, when, and how new terms can be added
- May or may not have structured relationships between terms
- To support indexing/tagging/metadata management of content to facilitate content management and retrieval

Definitions & Types

Specific types

- Term Lists/Pick lists
- Synonym Rings
- Authority Files
- Taxonomies
- Thesauri
- Ontologies

Definitions & Types: Specific Types

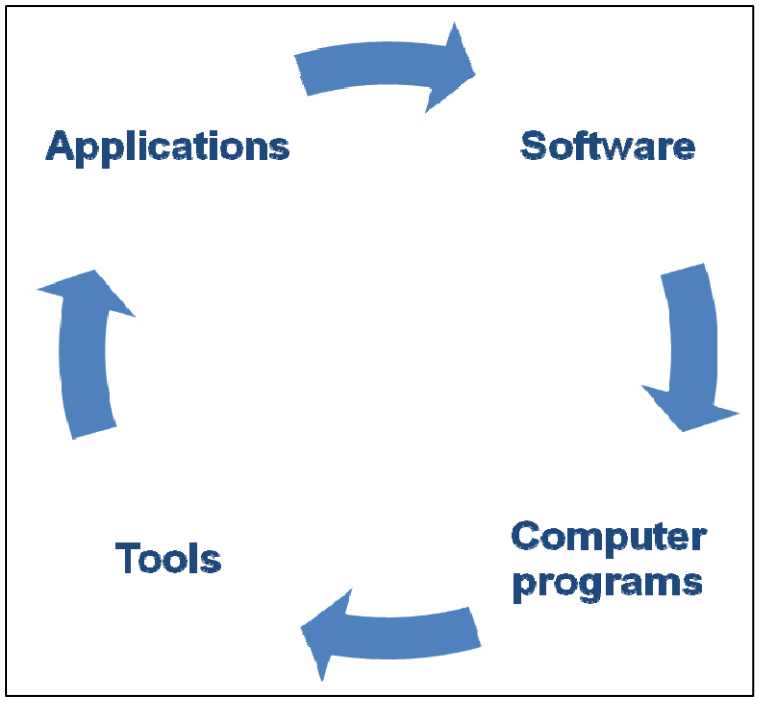
Term List

- A simple list of terms
- Lacking synonyms, usually short enough for browsing
- Often displayed in drop-down scroll boxes

Definitions & Types: Specific Types

Synonym Ring

- A controlled vocabulary with synonyms or near-synonyms for each concept
- No designated “preferred” term: All terms are equal and point to each other, as in a ring.
- Table for terms does *not* display to the user



	A	B	C	D
1	applications	software	computer programs	tools
2	administrative agencies	federal agencies	government agencies	
3	civil actions	civil litigation	civil cases	
4	agriculture	farming		
5	Americans with Disabilities Act	ADA		

Definitions & Types: Specific Types

Taxonomy

- A controlled vocabulary with internal structure.
- Terms are grouped or have hierarchical relationships.
- Emphasizes categories and classification for end-user display.
- May or may not have synonyms.

- Hierarchical – all terms have broader/narrower relationships to each other to form one big hierarchy
- Faceted – terms are grouped by attribute/aspect and are used in combination for indexing and search

Definitions & Types: Specific Types

Hierarchical Taxonomy example (UK's IPSV):

Top Level Headings

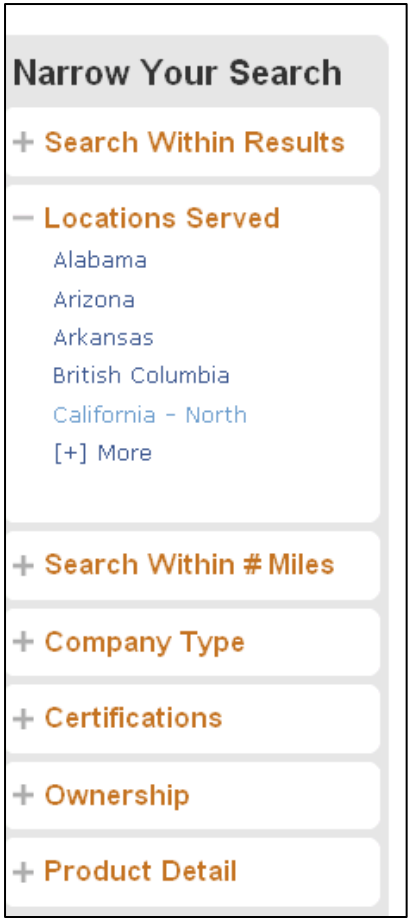
- Business and industry
- Economics and finance
- Education and skills
- Employment, jobs and careers
- Environment
- Government, politics and public administration
- Health, well-being and care
- Housing
- Information and communication
- International affairs and defence
- Leisure and culture
- Life in the community
- People and organisations
- Public order, justice and rights
- Science, technology and innovation
- Transport and infrastructure

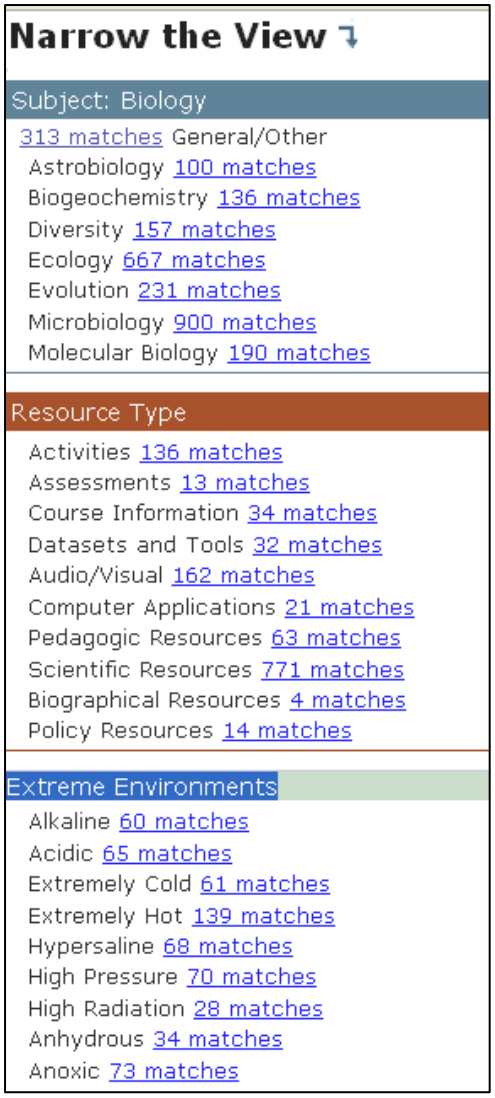
Leisure and culture

. Arts and entertainment venues
. . Museums and galleries
. Children's activities
. Culture and creativity
. . Architecture
. . Crafts
. . Heritage
. . Literature
. . Music
. . Performing arts
. . Visual arts
. Entertainment and events
. Gambling and lotteries
. Hobbies and interests
. Parks and gardens
. Sports and recreation
. . Team sports
. . . Cricket
. . . Football
. . . Rugby
. . Water sports
. . Winter sports
. Sports and recreation facilities
. Tourism
. . Passports and visas
. Young people's activities

Definitions & Types: Specific Types

Faceted Taxonomy examples





Outline

- Taxonomy Introduction: Definitions & Types
- Taxonomy Introduction: Purposes & Benefits
- Synonyms for Terms
- Auto-Indexing and Auto-Categorization
- Taxonomies for Auto-Categorization
- Taxonomy Resources

Purposes & Benefits

1. Controlled vocabulary aspect:
   Brings together different wordings (synonyms) for the same concept and disambiguates terms
   - Helps people search for information by different names
   - Helps people retrieve matching concepts, not just words

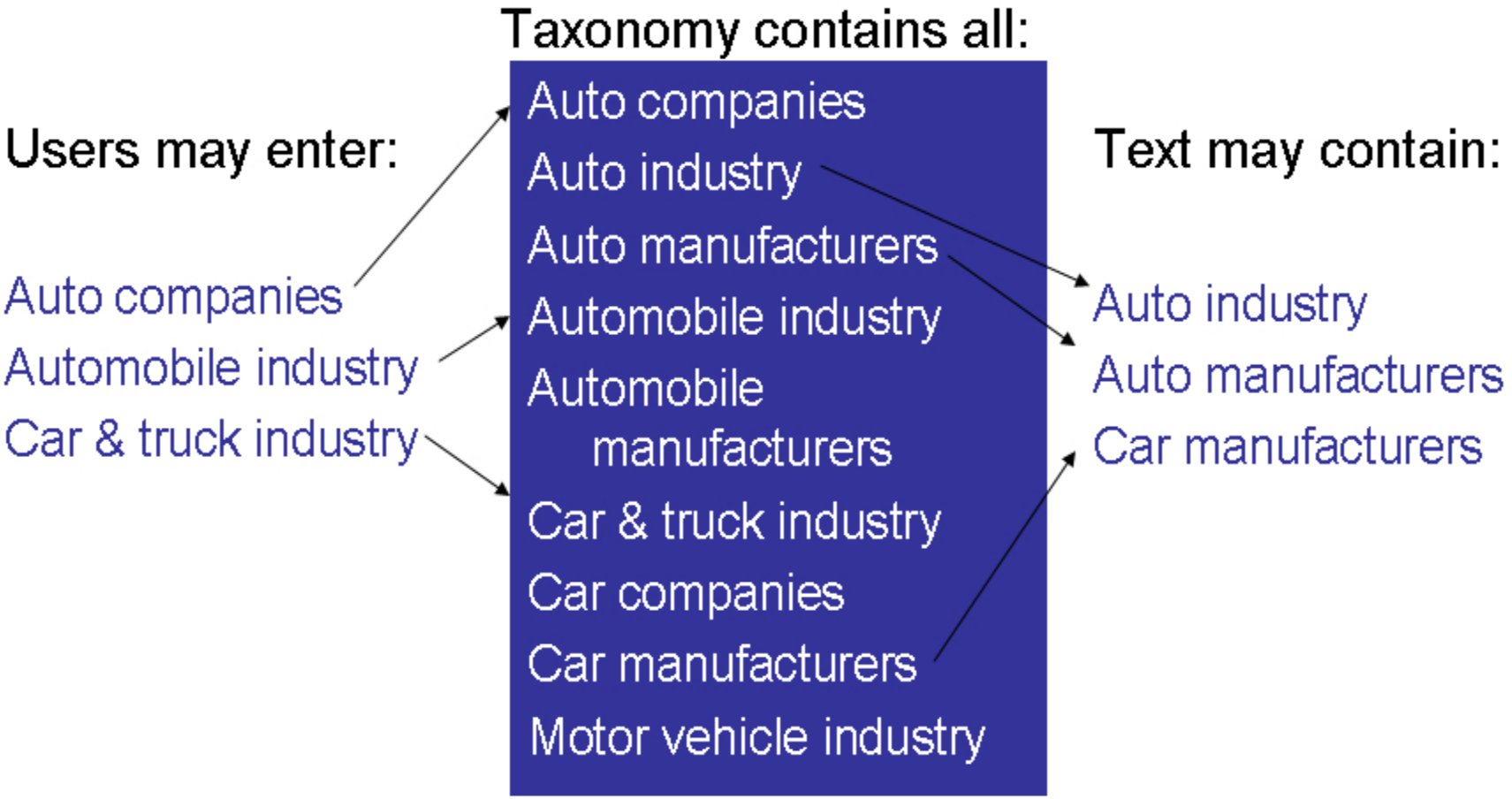
2. Taxonomy or thesaurus structure aspect:
   Organizes information into a logical structure
   - Helps people browse or navigate for information

Purposes & Benefits

- Helps people search for information by different names

- There are multiple ways to describe the same thing.
- A controlled vocabulary gathers synonyms, acronyms, variant spellings, etc.
- Without a controlled vocabulary keyword searches would miss some relevant documents, due to:
  - Use of different words (e.g. ***Attorneys***, instead of ***Lawyers***)
  - Use of different phrases (e.g. **Deceptive acts or practices** instead of **Unfair practices)**
  - User does not knowing the spelling of unusual names (e.g. ***Condoleezza Rice***)

Purposes & Benefits



Purposes & Benefits

- Helps people retrieve matching concepts, not just words

- A single term may have multiple meanings.
- Controlled vocabulary terms can be clarified/ disambiguated.
- Without a controlled vocabulary, too many irrelevant documents would be retrieved.
- A search restricted on the controlled vocabulary retrieves concepts not just words.
  - Excludes document with mere text-string matches (e.g. ***monitors*** for computers, not the verb “observes”)

Outline

- Taxonomy Introduction: Definitions & Types
- Taxonomy Introduction: Purposes & Benefits
- Synonyms for Terms
- Auto-Indexing and Auto-Categorization
- Taxonomies for Auto-Categorization
- Taxonomy Resources

Synonyms for Terms

- Supports search in most controlled vocabulary types: synonym rings, authority files, thesauri, (some taxonomies)
- Anticipating both:
  - varied user search string entries
  - varied forms in the text for the same content
- For both manual and automated indexing
- A concept may have any number of synonyms, but a synonym can point to only one preferred term
- Varied synonym sources:
  - Search analytics records
  - Interviews and use cases
  - Legacy print indexes
  - Obvious patterns (acronyms, phrase inversions, etc.)

Synonyms for Terms

Not all are “synonyms.”

Types include:

- synonyms: Cars USE Automobiles
- near-synonyms: Junior high USE Middle school
- variant spellings: Defence USE Defense
- lexical variants: Hair loss USE Baldness
- foreign language proper nouns: Luftwaffe USE German Air Force
- acronyms/spelled out forms: UN USE United Nations
- scientific/technical names: Neoplasms USE Cancer
- phrase variations (in print): Buses, school USE School buses
- antonyms: Misbehavior USE Behavior
- narrower terms: Alcoholism USE Substance abuse

Also called “variant terms,” “equivalence” terms, “non-preferred terms”

Outline

- Taxonomy Introduction: Definitions & Types
- Taxonomy Introduction: Purposes & Benefits
- Synonyms for Terms
- Auto-Indexing and Auto-Categorization
- Taxonomies for Auto-Categorization
- Taxonomy Resources

Auto-Indexing and Auto-Categorization

Choosing human vs. automated indexing:

Human indexing

- Manageable number of docs
- Higher accuracy in indexing
- May include non-text files
- Invest in people
- Low-tech: can build your own indexing UI
- Internal control

Automated indexing

- Very large number of docs
- Greater speed in indexing
- Text files only
- Invest in technology
- High-tech: must purchase auto-indexing software
- Software vendor relationship

Auto-Indexing and Auto-Categorization

Automated Indexing Technologies

- Entity extraction
- Text analytics and text mining, based on NLP
- Auto-categorization

Auto-Indexing and Auto-Categorization

Choosing auto-indexing methods:

Information extraction/text analytics

- For varied and undifferentiated document types
- For unstructured content
- For varied subject areas
- Terms may or may not be displayed
- Not necessarily with taxonomy

Auto-categorization

- For consistent doc types/formats
- For structured or pre-tagged content
- For limited/focused subject
- Displays categories to user
- Leverages a taxonomy

Combine both text analytics and auto-categorization:

1. Text analytics to extract concepts from unstructured varied content
2. Auto-categorization to apply benefits of a taxonomy/controlled vocabulary

Auto-Indexing and Auto-Categorization

auto-categorization = auto-classification = automated subject indexing

Auto-categorization makes use of the controlled vocabulary matched with extracted terms.

Primary auto-categorization technologies:

1. Machine-learning and training documents
2. Rules-based categorization

Auto-Indexing and Auto-Categorization

Machine-learning based auto-categorization:

- Automatically indexes based on previous examples
- Complex mathematical algorithms are created
- Taxonomist must then provide multiple representative sample documents for each CV term to “train” the system.
- Best if pre-indexed records exist (i.e. converting from human to automated indexing), then hundreds of varied documents can be used for each term.

Auto-Indexing and Auto-Categorization

Rules-based auto-categorization

- Taxonomist must write rules for each CV term
- Like advanced Boolean searching or regular expressions

Example:

Bush
IF (INITIAL CAPS AND (MENTIONS "president*" OR WITH administration*" OR
AROUND "white house" OR NEAR "george"))
USE
U.S. President
ELSE
USE Shrubs
ENDIF

Data Harmony

Outline

- Taxonomy Introduction: Definitions & Types
- Taxonomy Introduction: Purposes & Benefits
- Synonyms for Terms
- Auto-Indexing and Auto-Categorization
- Taxonomies for Auto-Categorization
- Taxonomy Resources

Taxonomies for Auto-Categorization

No matter which method of auto-indexing, auto-indexing impacts controlled vocabulary creation:

- Continual update work is needed (new training documents or new rules) for each new term created.
- Feeding training documents is easier for non-information professionals, than is writing rules

Taxonomies for Auto-Categorization

Taxonomies designed for auto-categorization:

- Need more, varied synonym/variant terms
- Need variant terms of different parts of speech
- Cannot have subtle differences between preferred terms
- Avoid creating many action-terms
- Taxonomy needs to be more content-tailored, content-based

Taxonomies for Auto-Categorization

Synonym/variant term differences:

For human-indexing

Presidential candidates

Candidates, presidential

For auto-categorization

Presidential candidate

Presidential candidacy

Candidate for president

Candidacy for president

Presidential hopeful

Running for president

Campaigning for president

Presidential nominee

Outline

- Taxonomy Introduction: Definitions & Types
- Taxonomy Introduction: Purposes & Benefits
- Synonyms for Terms
- Auto-Indexing and Auto-Categorization
- Taxonomies for Auto-Categorization
- Taxonomy Resources

Taxonomy Resources

- ANSI/NISO Z39.19 (2005) *Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies.* Bethesda, MD: NISO Press. www.niso.org
- Hedden, Heather. (2010) *The Accidental Taxonomist.* Medford, NJ: Information Today Inc. www.accidental-taxonomist.com
- American Society for Indexing: Taxonomies and Controlled Vocabularies Special Interest Group www.taxonomies-sig.org
- Special Libraries Association (SLA): Taxonomy Division http://wiki.sla.org/display/SLATAX
- Taxonomy Community of Practice discussion group http://finance.groups.yahoo.com/group/TaxoCoP
- "Taxonomies and Controlled Vocabularies“ Simmons College Graduate School of Library and Information Science Continuing Education Program, 5 weeks. $250. November 2012, January 2013. http://alanis.simmons.edu/ceweb/byinstructor.php#9

33

Questions/Contact

Heather Hedden
Hedden Information Management
Carlisle, MA
heather@hedden.net
978-467-5195
www.hedden-information.com
www.linkedin.com/in/hedden
twitter.com/hhedden
accidental-taxonomist.blogspot.com